***Горский Павел Владимирович***
*кандидат технических наук,*
*ведущий научный сотрудник сектора*
*программного обеспечения мониторинга РИЭПП.*
*Тел. 917-86-66,*
*info.riep.ru*

# ПОНЯТИЕ «БЛИЗОСТИ» В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ: ВЫЧИСЛИТЕЛЬНЫЕ ПРОБЛЕМЫ И ВОЗМОЖНЫЕ ПУТИ ИХ РЕШЕНИЯ

В контексте проблемы кластеризации предприятий под кластером понимается квазиинтегрированная структура, которая состоит из юридически независимых компаний, не располагает существенной рыночной властью, но в которой осуществляется контроль над управлением активами этих фирм. Другими словами, кластер может состоять из предприятий, специализированных в определенном секторе производства и локализованных географически.

Вместе с тем, ограничиваясь географическими рамками, мы изначально существенно сужаем возможности кластеризации. Развитие коммуникаций на всех уровнях, начиная от транспортного и заканчивая Интернетом, открывает перед возможным объединением предприятий новые перспективы. В связи с этим представляется интересным рассмотреть возможности кластеризации в рамках целой отрасли или даже группы смежных отраслей. Разумеется, ручными методами это сделать невозможно – придется привлекать автоматическую кластеризацию. При этом на первый план выходят проблемы, связанные с мерами близости (сходства), на основе которых автоматические методы работают [1].

Прежде всего, заметим, что кластеризация есть *неуправляемая* классификация, которую нужно отличать от *управляемой* классификации, каковым является дискриминантный анализ [2]. В этом смысле мы не можем заранее знать, сколько кластеров может быть выявлено. Поэтому удобно применять процедуры иерархической кластеризации, которые дают весь спектр (дерево) кластеров, а выбор того или иного уровня дерева остается за исследователем. Перед началом такой процедуры все объекты считаются отдельными кластерами, которые в ходе алгоритма объединяются. Вначале выбирается пара ближайших кластеров, которые объединяются в один кластер. В результате количество кластеров становится равным N-1. Процедура повторяется, пока все классы не объединятся. На любом этапе процесс объединения можно прервать, получив нужное число кластеров.

Отметим, что современные методы кластеризации могут работать как с количественными, так и с неколичественными данными. Неколичественные данные по своему происхождению – это, вообще говоря, данные, извлекаемые из текстовых документов, а следовательно, семантически плохо определенные; их структура не обязательно является регулярной. На фор-

мальном уровне единицей анализа является поименованная сущность (объект данных), описываемая произвольным набором элементарных свойств (качеств). Другими словами, сущность определяется как подмножество во множестве свойств / качеств. Свойство, в свою очередь, определяет, посредством своей встречаемости, группу сущностей, и, следовательно, может рассматриваться как подмножество во множестве сущностей. Таким образом, мы имеем симметрию, позволяющую обрабатывать сущность и ее свойства схожим образом: рассматривая набор данных как два множества, описываемых зависимостью типа «многие-ко-многим». Однако надо отметить, что, хотя такая симметрия не всегда осмыслена на уровне интерпретации, она всегда присутствует с формальной точки зрения. Поэтому возможный способ анализа существенным образом опирается на этот факт. На практике набор данных существует как последовательность записей, каждая из которых описывает один объект (определяет его имя и набор качеств). Качества могут принадлежать к различным группам. Эти группы могут служить аналогами переменных («полей» – в терминах баз данных), а качества, им принадлежащие, – значениям переменных. Но группы, с одной стороны, могут иметь более одного значения для каждой записи, а с другой стороны, их существование в общем случае необязательно. Более того, группы качеств могут существовать динамически и приобретать различный смысл в процессе анализа. Наша основная задача – определение близости между такими «группами качеств».

Рассмотрим основные меры близости, которые могут быть использованы при кластеризации предприятий. Большинство мер близости по существу являются мерами сходства или, напротив, несходства, либо могут быть сведены к ним. Наиболее известен так называемый «геометрический подход» к измерению близости (Эвклидова мера). Однако он не всегда приводит к наилучшим результатам, поскольку во многих случаях измерять следует не расстояние, а сходство (или несходство) между объектами.

Дадим следующие определения. Пусть пространство $R_+ = \{(x_1, \ldots, x_n): x_i > 0\}$.

Мерой *сходства* назовем функцию S: Q x Q -> $R_+$, обладающую следующими свойствами:

S1. $0 \le S(X, Y) \le 1$.

S2. $S(X, X) = 1$.

S3. $S(X, Y) = S(Y, X)$.

Иногда S2 заменяется более жестким требованием S2’. $S(X, X) = \max S(X, Y)$.

Симметричность меры сходства является не столь обязательным, сколь традационным требованием. Так, сходство предприятий X и Y нессиметрично, если в качестве его меры рассматривать поток компонентов производства из X в Y.

По аналогии с мерой сходства определим и меру *несходства*. Мерой несходства в пространстве $R_+$ называется функция D: Q x Q -> $R_+$, обладающая следующими свойствами:

D1. $0 \le D(X, Y) \le 1$.

D2. $D(X, X) = 0$.

D3. D (X, Y) = D (Y, X).

Очевидно, по заданной мере сходства S всегда можно построить меру несходства D=1–S и наоборот. Сходство объектов в принципе может быть измерено не только по наличию, но и по отсутствию у них одних и тех же признаков.

Обобщением понятия связи (корреляции) величин, измеренных в шкале не ниже порядковой, является понятие *соответствия*, которое может быть полным или частичным. Необходимое условие полного соответствия – равномощность сравниваемых множеств.

Во многих прикладных задачах разумно считать значения мер близости не «числовыми», а порядковыми. Это связано как с неточностью, «засоренностью» исходных данных, так и со специфическими характеристиками алгоритмов обработки данных. В этом смысле можно считать, что в большинстве случаев меры сходства и несходства, рассчитанные по описанию объектов некоторыми признаками, дают не количественную, а, в лучшем случае, порядковую информацию о близости. С другой стороны, для упорядочения объектов достаточно порядковой информации об их близости. Если считать, что близость объектов может измеряться в порядковой шкале, то равноценными можно считать любые меры близости, монотонно связанные друг с другом: ρ' и ρ'' эквивалентны, если для любых X, Y, Z, T

$\rho'(X, Y) \le \rho'(Z, T) <==> \rho''(X, Y) \le \rho''(Z, T)$.

Меры близости, удовлетворяющие такому условию, иногда называют комонотонными.

Интересно изучать сходства самих мер близости. Если при использовании каких-то мер близости в работе некоего алгоритма анализа данных всегда получаются достаточно близкие или сопоставимые результаты, то такие меры близости разумно считать сходными. Наиболее простой способ сравнения функций близости ρ' и ρ'' – это непосредственное сопоставление полученных с их помощью матриц близости $\| \rho' \|$ , $\| \rho'' \|$. Заключение об эквивалентности двух мер близости либо о наличии стохастической связи между ними может быть сформулировано в рамках проверки соответствующей гипотезы с помощью статистик, основанных на матрицах расстояний. Для сопоставления мер близости используются также методы многомерного шкалирования.

Характеризуя методы автоматической классификации с точки зрения возможности распространения выборочных результатов на генеральную совокупность, отметим, что статистические критерии значимости для проверки гипотезы о принадлежности объектов к тем или иным группам разработаны слабо. Полученная многомерная классификация может рассматриваться как характерная именно для изучаемой совокупности (как это обычно принято в анализе данных).

Особенностью представления параметров предприятий является тот факт, что множество параметров *E* может быть разложимо на на $k \le n$ групп признаков существенно различной природы, например, измеренных в различных шкалах: $E^i \cap E^j = \emptyset$ при $i \ne j$. Тогда существует k частных расстояний между *X* и *Y*, каждое из которых может быть определено с помощью неких частных мер близости. Все k частных расстояний можно считать из-

меренными в одной и той же шкале — шкале отношений, так что в принципе можно попытаться оценить среднее расстояние (если это не лишено содержательного смысла).

Нетрудно показать, что практически любые средние, используемые при обработке реальных данных о предприятиях, могут быть определены как решение оптимизационной задачи:

$$\sum_{i=1}^{m} \rho(X_i, X) \rightarrow opt \qquad (1),$$

где ρ — мера близости, а максимум (либо минимум) суммы близостей ищется на некотором множестве Q допустимых значений переменной X. Таким образом, теоретико-измерительные проблемы адекватности средних и мер близости тесно связаны [3].

Активное использование как самих эмпирических близостей, так и функций от них выдвигает ряд серьезных требований к выбору мер близости, адекватности алгоритмов анализа близости и обоснованности последующей интерпретации результатов. Исчерпывающее обсуждение всех вытекающих при этом вопросов вряд ли вообще возможно.

Обратим все же внимание на одну из наиболее актуальных проблем. Хорошо известно, что реальные данные далеко не всегда соответствуют той несложной модели, которую мы рассматривали выше.

Так, множество признаков *E* зачастую избыточно вследствие понятного опасения исследователя упустить некие факторы, относительную важность которых трудно определить заранее. С другой стороны, некоторые существенные признаки могут быть все же пропущены. Сказанное справедливо не только для множества признаков, но отчасти и для возможного множества рассматриваемых объектов. Наконец, в большинстве случаев доступные и включенные в рассмотрение признаки разнотипны – это и номинальные, и порядковые, и количественные переменные. Указанные обстоятельства подчеркивают стохастический характер исходных данных и приводят к выводу о необходимости статистической оценки измеряемого сходства, корреляции, расстояния и пр.

Для коэффициентов обычной и ранговой корреляции имеется развитое табличное обеспечение, позволяющее проверять гипотезу об отсутствии связи ($H_0$ : r = 0, где r – некий коэффициент корреляции) как при большом, так и при малом объеме выборки. Статистический анализ близости связан с существенно большими сложностями, прежде всего из-за затруднений с самой формулировкой нуль-гипотезы. Здесь можно предложить следующий подход. Поскольку набор номинальных признаков *E* выбирается априори, наблюдаемое совпадение их у объектов X, У может быть обусловлено случайными причинами. Нуль-гипотеза состоит в том, что распределение признаков на каждом из объектов – равномерное, а проверять ее можно, например, по отклонению величины мощности пересечения | X ∩ У | от математического ожидания (при $H_0$). Тогда, если мощность каждого из множеств X, У фиксирована, величина | X ∩ У | имеет гипергеометрическое распределение с параметрами | X |, | У |, n. Значимость величины

| X ∩ У | или, что то же самое, отклонения от нуля величины

| X ∩ У | - (1/n) | X |  | У | = 1/n (ab – bc) при заданном уровне доверительной вероятности может быть проверена с помощью соответствующих таблиц.

При описании предприятий нередко используют дихотомические признаки. Например, ответ на вопрос, есть ли у предприятия лицензия на производство данного вида продукции, может быть либо «да», либо «нет». Если мы рассматриваем пару предприятий, то дихотомическая мера близости между ними может быть описана в терминах четырехклеточной таблицы сопряженности:

– a – число признаков, отсутствующих у X и У одновременно,

– d – число совпадающих признаков,

– b (или c) – число признаков, присутствующих у X, но отсутствующих у У (или наоборот),

- a + b + c + d = n.

К настоящему моменту исследованы асимптотические свойства расстояний

D1 =  (b + c) / (a + b + c + d),

D4 = (b + c) / (b + c + d).

Найдены несмещенные оценки параметров асимптотического распределения вектора попарных расстояний, что позволяет строить соответствующие алгоритмы проверки согласованности, находить оценки для диаметра кластера и пр.

В случае разнотипных данных (k групп данных) целесообразно применять набор из k частных мер близости, каждая из которых инвариантна по отношению к допустимым в данной шкале преобразованиям.

Если веса в формулах ассоциативной меры близости порядка *p* не представляется возможным назначить априори, разумно будет перейти к ранжированным матрицам близости. В зависимости от результатов проверки конкордации k ранжировок попарных близостей следует либо построить среднюю в смысле (1) ранжировку ($H_0$ отклонена), либо провести кластеринг ранжировок ($H_0$ не отклонена) и привлечь дополнительную информацию для окончательного решения.

Рассматривая проблему измерения близости, мы убедились в том, что она не имеет простого и однозначного решения для всего многообразия задач кластеризации. В зависимости от существа поставленной задачи, характера и объема доступной информации и т. д. исследователь должен самостоятельно и последовательно проанализировать как теоретико-измерительные, так и содержательные аспекты этой проблемы в своем конкретном случае. Окончательный выбор может и не свестись к какой-либо одной море близости или к одному-единственному алгоритму анализа близостей. Однако разработанные к настоящему времени теоретические основы измерения близости позволяют существенно снизить область поиска и принять обоснованное решение.

Обратим еще внимание на проблему изучения результатов кластеризации, а именно – свойств кластеров. Одно из таких свойств – это плотность распределения точек внутри кластера. Насколько данный кластер является компактным, или же наоборот – достаточно разреженным. Несмотря на до-

статочную очевидность этого свойства, однозначного способа вычисления такого показателя (плотности) не существует. Наиболее удачным показателем, характеризующим компактность, плотность «упаковки» многомерных наблюдений в данном кластере, является дисперсия расстояния от центра кластера до отдельных точек кластера. Чем меньше дисперсия этого расстояния, тем ближе к центру кластера находятся объекты, тем больше плотность кластера. И наоборот, чем больше дисперсия расстояния, тем более разрежен данный кластер, и, следовательно, есть точки находящиеся как вблизи центра кластера, так и достаточно удаленные от центра кластера.

Необходимость обработки больших массивов данных приводит к формулированию требований, которым, по возможности, должен удовлетворять алгоритм кластеризации. Коснемся их кратко:

1) минимально возможное количество проходов по базе данных;

2) работа в ограниченном объеме оперативной памяти;

3) возможность прерывания работы алгоритма с сохранением промежуточных результатов, чтобы продолжить вычисления позже.

Алгоритм, удовлетворяющий данным требованиям (особенно второму), будем называть *масштабируемым*. Масштабируемость – важнейшее свойство алгоритма, зависящее от его вычислительной сложности и программной реализации. Имеется и более емкое определение. Алгоритм называют масштабируемым, если при неизменной емкости оперативной памяти с увеличением числа записей в базе данных время его работы растет линейно.

Однако вычислительные способности алгоритма – еще не гарантия успеха. Большую роль играет возможность учета специфики данных, которая часто не позволяет корректным образом использовать хорошо апробированные модели по нескольким причинам:

1) отсутствие представительной статистики, которую зачастую невозможно получить из-за больших материальных затрат на ее получение;

2) наличие пропусков в данных: восстановление или просеивание данных возможно при наличии представительной статистики;

3) интервальный характер данных, обусловленный неопределенностью условий их получения;

4) большая размерность признакового пространства, вызванная наличием нескольких десятков характеристик предприятий: известные методы сжатия «не работают», поскольку требуют проведения корректной нормировки.

Несколько слов стоит сказать и о пороговых значениях при объединении объектов в кластеры. Слишком низкая величина порога приводит к образованию чрезмерно большого числа близких классов, что замедляет как работу самого алгоритма, так и последующего этапа слияния. Слишком большая величина приводит к объединению в один кластер, например, существенно различных предприятий. Поскольку такая ошибка не может быть исправлена на этапе слияния, следует склоняться к выбору меньшего порога. Его численное значение может быть определено по критическим точкам распределения выбранной статистики критерия однородности.

Отметим также, что данные для кластеризации предприятий могут быть взяты в том числе из Геоинформационных систем (ГИС). Согласно теории

Л.А. Лютого [4], геоинформационные описания базируются на трех основных сферах представления знаний: невербальные знания, которые не могут быть представлены в вербальной форме; вербальные знания, которые не могут быть адекватно переведены в невербальную форму; и та часть знаний, которая может быть представлена в вербальной и невербальной формах. Современные алгоритмы кластеризации могут работать и с данными, представленными в невербальной форме (изображениями).

В заключение коснемся вопроса оценки эффективности кластеризации. Как отметил В.П. Третьяк, «с точки зрения отраслевого рынка, результативность функционирования малого бизнеса в кластере может оцениваться показателем доли малого бизнеса в выпуске отраслевой продукции. По некоторым международным нормам такой долей могла быть 30 % в отраслевом предложении. С точки зрения субъекта рынка, результативность функционирования малого бизнеса в кластере может оцениваться показателями самого кластера: прибыльность, восприимчивость инновациям, финансовые потоки и т. п. Кроме того, стремление войти в тот или иной кластер конкретной малой фирмы также можно рассматривать как показатель популярности кластера. В качестве показателей результативности функционирования кластера может выступать также наличие или отсутствие в нем третейских судов, общественных объединений, работающих на принципах саморегулирования, форм доверия между участниками кластера, прозрачности коммерческой информации внутри кластера» [5].

С этой точки зрения удачная кластеризация может в значительной степени содействовать эффективности кластеров предприятий.

## Литература

1. Раушенбах Г.В. Меры близости и сходства // Анализ нечисловой информации в социологических исследованиях. М.: Наука, 1985.
2. ACM Computing Surveys. Vol. 31. № 3. Sept. 1999.
3. Орлов А.И. Устойчивость в социально-экономических моделях. М., 1979.
4. Лютый Л.А. Язык карты: сущность, система, функции. М.: ГЕОС, 2002.
5. Третьяк В.П. Кластеры предприятий: пути создания и результативность функционирования // Интернет-конференция «Сетевые формы межфирменной кооперации: стратегические вызовы и конкурентные преимущества новых организаций XXI века», 2004.